Bridging the Gap Between Research and Practice: Predicting What Will Work Locally

By Kathryn E. Joyce and Nancy Cartwright

This essay addresses the gap between what works in research and what works in practice. Currently, research in evidence-based education policy and practice focuses on RCTs. These can support *causal ascriptions* ('It worked') but provide little basis for local effectiveness predictions ('It *will work* here') which are what matter for practice. We argue that moving from ascription to prediction by way of causal generalization ('It works') is unrealistic and urge focusing research efforts directly on how to build local effectiveness predictions. We outline various kinds of information that can improve predictions and encourage using methods better-equipped for acquiring that information. We compare our proposal with others advocating a better mix of methods, like 'implementation science', 'improvement science', and 'practice-based evidence'.

Keywords: causal claims, evidence-based education, educational research, RCTs, research-informed practice

For nearly two decades the dominant model for evidence-based education (EBE) has focused on improving schools by researching 'what works.' Yet, anyone familiar with EBE recognizes its relentless adversary: the gap between research and practice (Coburn & Stein, 2010; Farley-Ripple et al., 2018; McIntyre, 2005; Nelson & Campbell, 2017; Tseng & Nutley, 2014). The challenge is how educators can use research results to improve their outcomes in practice. Despite efforts to bridge the gap, primarily by more effective dissemination of results from 'high quality' experimental research, interventions adopted on the basis of recommendations often fail to be effective in practice. Many respond by rejecting EBE in its current form or by denouncing the entire enterprise (e.g. Archibald, 2015; Biesta, 2007, 2010; Hammersley, 2002, 2013; Smeyers & Dapaepe, 2007). Rather than opposing the dominant model, in which vast resources have been invested, we propose ways to shift and expand it to improve its performance. We begin with an analysis of why the gap exists and propose constructive advances to help bridge it. We argue that the research-practice gap reflects a gap between the causal claims supported by the experimental research results generally favored in EBE —'It *worked*'—and the causal claims that are relevant to practice—'It *will work* here.' Researchers may produce evidence to support the former, but those on the practice side

must figure out whether a program can work for them, and if so, what they need to put in place to get it to do so. But, so far, EBE has not been centrally concerned with producing and disseminating research that helps with this task. We conclude that addressing the gap requires a major rethinking of the research investigation and theory-building needed to support EBE and of the demands on research-users.

Drawing on guidance for deliberating in other policy areas, especially child protection, we provide a catalogue of some things that can help research-users – educators and decision-makers in their home sites—make more reliable predictions about what might work, and how it might do so, for their school, their district, their students. Discussing the knowledge-*use* side of EBE carries implications for organizing the knowledge-*production* side and for knowledge-mobilization. Our principal contribution on the knowledge-production side is in diagnosing the gap and showing that bridging it requires dramatically expanding the kinds of evidence that are collected and disseminated by EBE and adjusting the methods used to judge its acceptability.

EBE relies on researchers to produce evidence of effectiveness for educators to use in practice. Intermediary organizations like the What Works Clearinghouse (WWC) in the U.S., the What Works Network and Educational Endowment Foundation in the U.K., the European EIPPEE Network (Evidence Informed Policy and Practice in Education in Europe) are supposed to help bridge the gap between research and practice by evaluating research, summarizing results, and advertising interventions that have proven efficacious in rigorous experimental studies, especially randomized controlled trials (RCTs).

Proponents of EBE commonly attribute the gap between outcomes in research and practice to deficiencies in how tasks are performed on one or both sides of the knowledge-production/ knowledge-use divide. Accordingly, plans for addressing it encourage researchers to conduct more relevant research, offer guidance for implementation (Gutiérrez & Penuel, 2014), translate "research-

based knowledge into…generalized practical suggestions" (McIntyre, 2005, p. 364), and effectively communicate research findings to decision-makers in easily digestible formats and in ways that encourage use. For example, Levin (2013) explains that research findings must compete with information from other sources that influence potential users. He suggests using knowledge mobilization strategies that appreciate how experience, organizational practices, and attitudes shape educators' engagement with research. On the practice, or knowledge-use, side, strategies generally emphasize the importance of cultivating evidence-literacy and exercising professional judgment when choosing and implementing interventions (Brown & Schildkamp & Hubers, 2017; Bryk, 2015). By contrast, we are concerned with the *information and reasoning* needed to bridge the gap between the kind of causal claim supported by research -- *causal ascriptions* -- and the kind of claims that are relevant to practice -- *effectiveness generalizations* and *effectiveness predictions.*

RCT results, along with meta-analyses and systematic reviews of them, are taken across evidence-based policy communities as 'gold-standard' evidence for 'what works'. There is much debate about this concentration on RCTs (see section I below). Our concerns are with sloppy talk. Currently, EBE is plagued by casual use of language that is imprecise about what kinds of causal claims are at stake and also about what kind of work it takes to warrant the kinds needed in practice. We will look carefully and critically at just what kinds of claims can (in the ideal) be warranted by RCTs and contrast that with the kinds of claims educators need to know. Then we will turn to *the argument theory of evidence* and the related *material theory of induction* to lay out what it takes, beyond the much discussed RCT, to provide evidence for claims that a policy 'works' generally or that it will work *here.*

We shall defend the claim that RCTs can provide evidence for *causal ascriptions*— the intervention *worked* in the study population in the study setting. But, as the argument theory shows, it takes a great deal of additional knowledge besides an RCT result to warrant the claims educators

need – *effectiveness generalizations* and *effectiveness predictions.* Yet, the WWC, like other organizations driving EBE, claim that "high-quality research" from RCTs can show "what works in education" and support "evidence-based decisions" (WWC, n.d.). Similarly, Connolly, et al. (2017), assert that "RCTs within education offer the possibility of developing a cumulative body of knowledge and an evidence base around the effectiveness of different practices, programmes, and policies" (p.11). Like other philanthropic organizations involved in social policy, the Gates Foundation calls for schools to use methods "grounded in data and evidence." Their programs focus on "identifying new, effective approaches that can be replicated in other schools" and "using evidence-based interventions and data-driven approaches to support continuous learning" (Gates Foundation, n.d.).

As these statements indicate, EBE aims to support predictions concerning how an evidence-based intervention will perform in a new educational setting *indirectly* by establishing effectiveness *generalizations*. While generalizations would naturally justify predictions about specific cases, we argue that they are unnecessary. This is a good thing since there are few useful, reliable general effectiveness claims to be had. The kinds of claims that can be expected to hold widely in education, or even over restricted domains characterized by a handful of descriptors (inner-city, free-school lunches, ESL, high-achievers, etc.), are usually too abstract to guide practice. For example: Children learn better when they are well-nourished, have a secure environment, read at home, and have adequate health care.

Instead of just trying to support local plans and predictions indirectly by establishing 'what works,' we argue that EBE should focus on research that supports these *directly* by producing evidence research-users can employ to make effectiveness predictions locally. We should note at the start that because context does significantly affect effectiveness, it will seldom be possible to replicate results by moving a program as implemented in study sites to new settings. If the program is to work in a context it will have to be *fitted* to that context, and this seldom involves just tinkering

around the edges or implementing it well. Rather it requires getting all the right features in place that will allow the program to work *here* and guarding against features that can derail it *here*. So, the work is harder than one might hope: If the program is to be effective, a context-local program-plan must be built. And to build a program-plan that you can expect to be effective for the goals you want to achieve in a specific context, you need to know what facts make it likely that the program will work *here*. Since just what these facts are can vary from context to context, often dramatically, there's no telling in advance what they will be. But we can tell what *kinds* of facts matter. To this end, we will outline a variety of kinds of information that can guide local planning and make local predictions about the effectiveness of proposed plans more reliable and more useful.

The kinds of research that can produce the requisite information, locally or more generally, often in co-production, requires a mix of methods well beyond those listed in current evidence hierarchies. This suggestion resonates with other recent calls for more mixed-method research in education,[1] but our reasons are different from the usual. The standard reasons for mixing methods in EBE are to aid implementation (Gorard et al., 2017) and to make general effectiveness claims more reliable (Connolly et al., 2017; Bryk, 2015). We, by contrast, encourage mixed methods because reliable and useful effectiveness predictions require a variety of different kinds of information relevant to determining how an intervention will perform in a specific setting that different kinds of research help to uncover. While local effectiveness predictions will never be certain, incorporating this information can improve them.

Section I situates our project within the general intellectual setting and clarifies the contribution we wish to make. Section II provides an account of how the current research emphases in EBE contribute to the research-practice gap. Building on a framework[2] that distinguishes different kinds of causal claims prevalent in evidence-based policy, we assess the role that RCTs can play for each, underlining that RCTs are poor evidence for what educators need to know: Will this

intervention work here in our school or classroom, for these students? Section III offers a catalogue of kinds of information that can help educators make local predictions that are reliable and useful. We urge developing and using a broader panoply of research methods than are generally endorsed in EBE to help uncover these kinds of information. Section IV compares other approaches that call for a better, broader mix of methods, like 'implementation science', 'improvement science', and 'practice-based evidence' with our own.

I. Framing Our Project

We acknowledge that debates about education research methodologies often stem from fundamental disagreements about the epistemology and ontology underpinning various approaches (e.g. Bridges, 2017; Bridges & Watts, 2008; Crotty, 1998; Howe, 2009; Phillips, 2007; Smeyers & Depaepe, 2007). In particular, many critics of EBE target the positivist approach they take it to represent. Our project does not contribute to these debates. Disagreements about the meaning of broad positions like 'positivism' abound; yet, differences between the relevant ontological and epistemological positions are difficult to pin down, as are their precise consequences for debates concerning EBE. Additionally, interlocutors on all sides disagree about whether the disputed positions bear a *necessary* relation to EBE.[3] Although they continue to favor RCTs for investigating causal questions, many education researchers involved in EBE now take their background theories to be compatible with a wide range of methodologies (e.g. Connolly et al., 2017; Gorard et al., 2017). Instead of attempting to navigate this familiar, fraught landscape, we articulate specific problems underlying the research-practice gap without relying on controversial terminology or attempting to identify their ideological origins. Still, readers will find that our arguments reflect and address some of the prevalent criticisms of EBE driving these debates—namely, the tendency to neglect the importance of context, of what teachers, students, and parents can contribute, and of professional

judgment (Biesta, 2007; Smeyers & Depaepe, 2007).

However one classifies the dominant EBE model, it is part of a larger evidence-based movement across areas of social policy and is enshrined in U.S. federal education policy (Eisenhart & Town, 2003; Education Sciences Reform Act, 2002; No Child Left Behind Act, 2002; Every Student Succeeds Act, 2015). We start from their assumption that research can help improve students' educational outcomes and experiences and hunt for ways to build and improve on the widespread efforts in this direction that are already in place. As we understand it, this assumption need not imply that social phenomena yield to the same study methods as natural phenomena, nor does it entail an instrumentalized view of education, a narrow conception of causality as deterministic, or denying that education involves agents who engage in socially-developed normative and cultural processes (Bridges & Watts, 2008; Biesta, 2007, 2010; Smeyers & Depaepe, 2007). To the contrary, we hold that effects issue from a plurality of context-local causal factors that contribute to them, including the actions, attitudes, norms, and habits of students, teachers, parents, school administration, and the wider neighborhoods in which these are embedded. Because we often cannot identify, untangle, or measure all of the factors that contribute to an outcome, our causal claims are never certain. Still, the warrant for them can be better or worse.

We recognize that assessing effectiveness is only one part of a larger decision-making process that involves considering values, setting goals, and mapping local assets (Brighouse et al., 2018). Educational policies and practices should be fair, compassionate, and effective; and it is difficult to predict if any of these, let alone all three, will be true for a policy we consider using *here*, as we are planning to implement it *here*, given the complicated set of interacting factors that are relevant *here*. We cannot tackle all three in this paper. We offer positive suggestions on the third – effectiveness – which is EBE's central concern. Expectations can always go wrong, both ethically and epistemically, but care in deliberation can improve predictions on both sides. Our suggestions

here focus on the epistemic side, more specifically, on whether policies as planned for the local context will achieve their targeted aims, setting aside important ethical questions regarding choice of target and who is advantaged and disadvantaged by those choices.

Because we situate our arguments within the existing EBE framework, they ought to be of interest to proponents seeking to improve its performance. At the same time, our diagnosis captures some of the familiar concerns commonly raised by anti-positivist critics. Like them, we reject the intervention-centered approach to EBE that focuses exclusively on 'what works.' Instead, we endorse a context-centered approach that starts from local problems in local settings with their own local values, aims, resources and capabilities to provide tools that help predict what will work there and what students, teachers, parents, and school staff can contribute to success.[4] Framing our project this way allows readers with differing commitments to seriously consider and even accept our arguments, albeit they may do so for different reasons.

II. Diagnosing the Problem

II. 1. Getting clear on just what claims are being made

Clarifying our concept of evidence helps distinguish distinct causal claims, which are often conflated without note to bad effect. Evidence is always evidence for (or against) *something*. The question is not 'What makes a result good evidence?' but rather 'What makes a result good evidence for *a particular claim*?' Using the concept of *evidence* imprecisely within EBE generates confusion that can undercut its success (Joyce & Cartwright, 2018; Kvernbekk, 2016; Spillane & Miele, 2007). A fact counts as evidence for a specified claim when it speaks to the truth of that claim. The *argument theory of evidence* provides a good way to capture this idea (Cartwright, 2010, 2013, 2017; Reiss, 2013; Scriven, 1994). According to the argument theory, something is evidence for a claim when it serves as a premise in a sound argument for that claim. Sound arguments are comprised of trustworthy

premises that jointly imply the conclusion. To be trustworthy, each premise must be supported by good reasons. These can include, among other things, empirical facts established by research, observations, and credible theory.[5]

The argument theory is a close cousin of the philosopher John Norton's highly regarded *material theory of induction*. Norton (2003) points out that there are serious problems with all the standard attempts to articulate a theory of inductive inference that depend on form alone (as in the case of induction by simple enumeration, which we discuss later). What really does the work, he maintains, are material facts, encoded in substantive claims that connect the evidence with the hypothesis to be evidenced, showing just *why* the putative evidence is evidence for that hypothesis. For this theory too, a research result is evidence *relative to* a target hypothesis and to a set of additional claims describing material facts about the world, including often general truths.

By contrast, some scholars use 'research findings' or 'research,' 'knowledge,' and 'evidence' interchangeably (e.g. Brown, 2014; Nutley & Walters & Davies, 2007). Their usage implies that when a study meets the espoused standards, its results count as 'evidence'. This does not make sense. We must know what the claim is to decide whether—or under what further assumptions—some finding counts as evidence for it, and how. Labeling research results as 'evidence' obfuscates differences in the claims for which EBE needs evidence.

II.2. What kind of claim can RCTs support?

Much of the literature advocating RCTs stresses the benefits of randomly assigning subjects to intervention and control groups in order to balance the other causal factors that might affect the outcome other than the intervention (e.g. Borman, 2009; Connolly et al., 2017; Gorard et al., 2017; Shavelson & Towne, 2002; Slavin, 2002). There are two problems with that.

First, even if randomization did achieve balance, it only balances other factors at the time of

selection. Correlation with other causal factors can arise post-randomization. For example, getting new materials may encourage effort or confidence that aids performance among students in the intervention group. Where possible, experiments use blinding to reduce this problem. Ideally, research subjects, those delivering interventions, those measuring outcomes, and those doing the statistical analysis, are all unaware of who received which treatment. But, at best, blinding works poorly in educational experiments. Teachers, students, and administrators know who is receiving the intervention and who is in the comparison group. This knowledge alone can impact outcomes, as can the training and support for teachers delivering the intervention.

Second, even supposing random assignment and successful blinding, we should not expect to get two groups that are the same with respect to all causal factors affecting the outcome other than the intervention. Getting balance like that in a study of, say, 100 units is like getting 50 heads in 100 flips of a fair coin. Of course, generally, the larger the sample the closer the results on a single run can be expected to be to the true average. This is one of the reasons why the WWC encourages multiple, large RCTs.

What we can expect if randomization, blinding, etc. are successful is this:

> *RCT conclusion:* The measured outcome – the difference between the observed mean in the treatment group and that in the control group – gives
>
> a. the average treatment effect[6]
> b. of the intervention provided in the treatment wing of the study *compared to* that provided in the control wing
> c. *in expectation* – across a hypothetical infinite sequence of runs of the experiment with a new randomization on each run
> d. for the population enrolled in the study.

We need to be attentive to each of these:

a. We only learn average results. Individuals in the study will generally have differed in their responses to the intervention and some may even have been harmed by it.

b. Exactly what the control group received matters significantly to the size of the measured treatment effect – the effect will seem much bigger if the control group received an intervention that performs badly than if it received one that performs well.

c. The estimate is, in technical language, *unbiased*. This has nothing to do with how close it is to the truth.[7]

d. The conclusion can at best be a causal *ascription*.

It is this last that we want to discuss in some detail.

Studies – any study, RCT or otherwise – can only support results about the things studied. Conclusions about things not studied must depend on assumptions outside the range of the study. To play the role EBE assigns them, RCTs would have to be almost sufficient by themselves for effectiveness claims. But what a positive result from a very well-conducted RCT can directly support is just a claim of the form we have labeled 'RCT conclusion', which is about the study population. Indeed, the WWC explicitly states that its standards "focus on the causal validity within the study sample (*internal* validity)" (WWC, 2017a, p. 1 original emphasis). It should be obvious that a causal ascription in a study population cannot directly evidence a general effectiveness claim or an effectiveness prediction. The fact that an intervention *worked* somewhere cannot show that the intervention *works* or that it *will work* in some other target.

II.3. Using results outside the study population

Assume for the moment that the WWC succeeds in vetting RCTs for what is called 'internal validity', i.e. to ensure the RCTs really do support causal ascriptions in the study populations. Educators are concerned with whether the tested intervention can positively affect outcomes in *their*

setting. The central question for them is not what is usually called 'external validity': Will the *same results* hold here? Generally, the answer to that is 'probably not' given that the local factors that determine an effect size vary from setting to setting. Rather, the central question is '*What* must be in place here for the intervention to deliver good enough results for us?' And then, 'Are the gains worth the expected costs?' For guidance, the WWC combines study results that meet standards for internal validity to provide an effectiveness rating and an "improvement index" for each intervention. In both cases, it explicitly uses evidence supporting causal ascriptions and effect sizes within study groups as evidence for general effectiveness claims. For example, the WWC rates *Read 180,* a multi-faceted literacy program, as having a positive effect on comprehension. The effectiveness rating rates the strength of evidence supporting the claim that *Read 180* positively affected outcomes within a domain (e.g. comprehension) based on "the quality of research, the statistical significance of findings, the magnitude of findings, and the consistency of findings across studies" (WWC, 2016a). Interventions, including *Read 180*, merit the highest effectiveness rating when these factors provide "strong evidence that an intervention *had* a positive effect on outcomes" (WWC, n.d., our emphasis).

The WWC calculates the *effect size* for each individual study by dividing the observed average treatment effect (the observed difference in mean outcomes between treatment and control) by the standard deviation of the outcome in the pooled treatment and control groups (WWC, 2017b). This means that all studies are recorded in 'standard deviation units', which, it is hoped, provides some sensible way to compare sizes for studies that use different scales for outcomes. Then, it averages those (standardized) effect sizes to produce the effect size for the domain. For *Read 180*, the effect size, averaging across all six qualifying studies, is 0.15. This average provides the basis for the improvement index, which depicts "the size of the effect from using the intervention" in a way that is supposed to help educators "judge the practical importance of an intervention's effect" (WWC,

2017b, p. 14). That is, it is supposed to represent the expected change in outcome for "an average comparison group student if the student *had* received the intervention" (p. 14, our emphasis). Based on expected improvement for an average comparison group student, the improvement index for *Read 180* is +6. Explaining its significance, the intervention report for *Read 180* says that if given the intervention, the average student "*can"* be expected to improve their percentile rank by six percent (WWC, 2016b).

The presentation of these findings implies that the stronger the evidence for causal ascriptions (represented by effectiveness ratings), the more credible is a general effectiveness claim or an effectiveness prediction. Educators are thus led to believe that 'strong' evidence and large effect sizes indicate that the intervention will be highly effective, producing significant effects for them.[8] For instance, because *Read 180* earns the highest effectiveness rating, educators can reasonably expect their students to improve by six percent, on average. Apparently, using RCTs justifies the inference from internal validity in each study and consistent results across multiple studies (or faring well in a meta-analysis of study results) to this conclusion about what can be expected in general.

Clearly this is a mistake. It is only the consistency across studies that is any indicator that the result might hold generally or at my school for my students; and without more ado – much more ado – it is a weak indicator at that. Adding up causal ascriptions as the EBE literature seems to recommend amounts to induction by simple enumeration, which has long been condemned as a weak form of inference.[9] When it *does* work, as it sometimes may in education, it is because the feature being generalized is projectable across the domain in question—*not* because the inference is based on *multiple* RCTs or combined findings from multiple RCTs. Any education examples we cite will likely prove controversial, so instead to make the point clear we illustrate with the example of electric charge: A negative charge of 1.602177×10^{-19} Coulombs measured on a single electron is

projectable across all electrons because we have strong reasons from theory and a multitude of different kinds of empirical results that all electrons have the same charge. The same form of inference yields a false conclusion if it generalizes a feature that does *not* project onto the target population. Consider the oft-cited case of the color of swans. Multiple samples consisting of only swans in London's Regents Park do not license the inference that all swans are white. Likewise, even if multiple RCTs show positive results, predictions need further premises showing that the causal relations being generalized to other students or schools projects onto them, as in the case of electrons. Without such premises, we risk drawing false conclusions, as with the swans.

What kind of reasons can support projectability in educational contexts? Imagine that a study has produced a good estimate of an average treatment effect in the study population and you want to argue that that estimate should hold for some target population. This would be warranted if you could argue that the students in the study sample are 'representative' of the target. One way to warrant that is to draw those students randomly from the target, where the target could either be a broad range of U.S. students or the students in a particular district or school. If the study group genuinely *represents* the target, then whatever probabilities are true of the study population (like the average effect of an intervention) are true of the target– that's what it means to be representative. But of course, just as random assignment should not be expected to produce balance of other factors in any one run of an experiment, random sampling should not be expected to produce a representative study population on any one draw of a sample to study. The probabilities of the study sample and the parent only match in expectation—over a hypothetical infinite series of trials. For any one study, the average treatment effect may be accurate for the study group but far off from the parent.[10]

Even without this worry, within education, transferring results from study to target population usually cannot be justified on these grounds because samples are seldom randomly

selected from the target population. Sometimes though a sample is drawn randomly from the students in a specific school or district, often with the plan to 'scale up' the intervention to the whole school or district if the study results are positive. But even if you draw your study students randomly, there are notorious difficulties in scaling up – the study population may not be representative of the whole population since people behave differently when a few receive the treatment than when all do.

Nor is it easier to construct study populations that are representative of a broad range of U.S. students. Researchers work within constraints that impose criteria for selecting study populations. For instance, researchers might be confined to a school district. Within the district, they must conduct the study within schools that are willing and able to participate. Even if all schools in the district are obliged to participate if randomly selected, that district becomes the parent population. The district was not randomly selected, so the experiment cannot support generalizing average effect sizes beyond the district.

A second problem is that random sampling from the target is not enough to ensure that the study results are representative of the target facts since much can happen to differentiate the two after sampling. Just as in an RCT, what happens to the sample after it is drawn may change the probabilities for relevant factors so that the sample is no longer representative of the parent. For instance, monitoring both groups may impact student performance independently of the intervention itself, which inflates effect sizes. Moreover, educational contexts are dynamic systems (Reynolds et al., 2014). Changes may occur that have no connection to implementing the intervention but still impact the reliability and generalizability of results.

The alternative to random sampling from the target is to argue that the target and the study populations are the same, or similar enough, in the ways that affect the results to be projected (e.g. the size or the direction of the effect). This indeed can be – and often is – the case. Then the result

from the study is evidence for the conclusion drawn about the target. But as the argument theory demands, it is only evidence relative to the additional assumption that the two populations are alike in the relevant respects. For example, if students' background knowledge or out-of-school resources impact the intervention's performance in the study sample, then students in the target population must have relevantly similar background knowledge or resources. For *credible* evidence-based policy or practice, the assumption that populations are alike in the relevant ways itself needs to be backed up by good reasons (see Joyce, 2019), which can come from both theory and other empirical results.

This is important in education because, as we have repeatedly remarked, it cannot be taken for granted that differences between educational contexts are negligible. Nuthall (2004) observes that "teaching one specific class of students is different from teaching any other class of students" because "engaging the students in relevant learning activities requires a unique understanding of each student's interests and relationships with classroom peers" (p. 294). Thus, Nuthall concludes that "what works in one classroom or with one student will not necessarily work in other contexts" (p. 294). Nuthall's insight that students within and across educational contexts have different needs to which educators must be responsive in practice seems especially important in contexts of inequality, where there are significant disparities in resources among students, schools, or districts.

Research on school effectiveness and improvement conducted over the past few decades demonstrates the complex, dynamic nature of learning environments (e.g. Reynolds et al., 2014). Researchers in these fields argue that various factors within schools and broader education systems and outside of school impact students' learning outcomes (Sarason, 1990; Lareau, 2011; McLeod, 2004). For example, such research highlights the importance of school culture (Freiberg, 1999; Hargreaves, 1995; MacNeil et al., 2009; Sarason, 1996). As Halsall (1998) observes, "one of the most consistent messages from the school improvement literatures is that school culture has a powerful impact on any change effort" (p. 29). Those who study its impact aim for "understandings of

sociocultural or organizational factors at the school level that facilitate or impede school improvement" (Schoen & Teddlie, 2008, p. 148). This literature indicates that even schools that share broad superficial characteristics like population density, range of socio-economic statuses, or urbanicity may differ in ways that bear on an intervention's effectiveness. Regardless of whether or not one accepts conclusions from school effectiveness or improvement research, the assumption that differences beyond superficial characteristics are irrelevant or inconsequential for assessing representation is a big assumption that requires serious justification.[11]

To judge when study and target settings are similar enough in the right ways takes *theory* – lots of it and of very different kinds. To learn just what it is about two different settings that allows them to support similar causal pathways from intervention to outcome requires a wealth of knowledge. This includes not only the theory of the process by which that intervention is supposed to bring about the intended result (sometimes called 'the logic model' or the 'theory of change' of the intervention) but also what contextual factors in the setting can support that process – for example, what are the psychological and sociological mechanisms at work and what can we expect them to do in this setting, what kinds of things in the economic and material structure can facilitate and what kind can hinder the process, and how can these be expected to interact in the target setting? None of this is to be learned from RCTs, even bigger and better ones. Still, it seems that conducting more large-scale RCTs remains the primary EBE strategy for improving evidence for effectiveness claims.

There is good reason to think that focusing narrowly on RCT results without considering contextual factors has sometimes encouraged harmful consequences, especially for students from marginalized groups. Helen Ladd (2012) argues that EBE and associated reform strategies (e.g. test-based accountability systems) are potentially harmful because they "pay little attention to meeting the social needs of disadvantaged children" (p. 204). Attempts to improve school quality ignore the

contextual factors associated with poverty "that directly impede student learning" (p. 219; see also Duncan & Murnane, 2011, 2014). One particular harm concerns the unequal distribution of highly credentialed teachers. The idea that interventions work, when implemented with fidelity, de-emphasizes the importance of quality teachers by suggesting that disparities in teachers' abilities across schools are not a problem as long as those in schools serving disadvantaged students are good enough to implement effective interventions faithfully.

Another harm stems from the fact that "instructionally focused interventions pay insufficient attention" to students' "socioemotional-learning needs" (Rowan, 2011, p. 534). Brian Rowan points out that benefitting from even high-quality instruction usually requires certain socioemotional skills, which correspond to social advantage (Becker and Luthar, 2002; McLeod & Kaiser, 2004; Rowan, 2011). Failure to recognize this can conceal the need to provide students with socioemotional learning opportunities and support. At the same time, overestimating the impact of students' behavior or social skills on learning (which seems sometimes to stem from implicit biases) can lead to overemphasizing discipline (e.g. some high-commitment 'No Excuses' charter schools) to the detriment of socially disadvantaged students (see Brighouse & Schouten, 2011; Curto et al., 2011; Tough, 2013). Further, thinking only about the effectiveness of interventions risks neglecting the quality of students' experiences, which can bear on outcomes but are also important in their own right. While here we focus on the general question of how to bridge the research-practice gap, these unintended side-effects underscore the importance of shifting toward a context-centered approach to EBE.

III. Bridging the Gap

For EBE to help improve educational outcomes, it must generate evidence that supports effectiveness predictions: 'This intervention will work here, as we plan to use it.' One indirect route

to this is via establishing a *general* effectiveness claim: 'If it almost always works, within a broad range of contexts and ways of implementing it, it will probably work here.' But, establishing these is a demanding task. Supporting such claims requires showing that an intervention has a *stable causal capacity*. For example, aspirin has a relatively stable capacity to relieve headaches: It reliably does so across wide-ranging circumstances and populations. It is unclear whether educational interventions can have stable capacities and, if at least some of them do, how we could arrive at them given the nature of educational contexts. As Berliner (2002) remarks, educational researchers conduct their investigations "under conditions that physical scientists find intolerable" (p. 19). However, establishing general effectiveness is not necessary for making reliable predictions of the sort educators need.

We suggest less emphasis on general effectiveness claims, which are hard to come by, in favor of research that produces evidence educators can use to make causal predictions locally. After all, for educators, a successful intervention is one that contributes to a positive effect for them regardless of whether it can do so elsewhere. Incorporating a prediction phase into the EBE model means educators and policymakers must do more than choose and implement interventions that are deemed to work. To plan policy and predict effectiveness, they must construct arguments with multiple premises supported by various types of information, which will necessarily involve theory as well as empirical results of various kinds. Because their arguments require complex local knowledge and judgment, educators themselves can have much of the information needed to make predictions for their settings. But most premises need to be warranted by other types of evidence, some of which researchers can supply. Namely, researchers can provide information to help educators recognize just what facts may be relevant for predictions regarding particular interventions. And many of these facts will be theoretical in nature. It is no good just piling up study results. If we want research to be useful to practice there is no way to avoid the heavy work of

detailing the theory of change of the intervention, of developing new concepts like metacognition, meta-affective awareness, stereotype threat, Bloom's Taxonomy (and its revision), and of generating and vetting general claims and mechanisms well beyond those of the form 'It works'.

We can take a lead here from recent work on evidence and child protection. Munro et al. (2016) outline three kinds of information that contribute to reliable local effectiveness predictions about an intervention:

1. Knowledge that the local social, economic, cultural, and physical *structure* can afford the necessary causal pathways for the intervention to lead to the outcome
2. Knowledge of what the *support factors* are for that intervention to work in the local setting and of whether these are available or can be made available
3. Knowledge of what *derailers* might interfere along the way to diminish or entirely halt progress toward the outcome.

Structure. Can the intervention work here or do local conditions simply not afford the steps that it takes, start-to-finish, (which are hopefully outlined in the theory of change of the intervention) to produce the outcome? For instance, no amount of tinkering with the reward structures for parents within resource constraints will get more children vaccinated in places where there are no vaccination clinics within reach and there is a strong cultural resistance to vaccination. Similarly, creating systems to incentivize or hold teachers accountable for students' performance is unlikely to improve outcomes at schools where teachers are already doing their best and low-performance is primarily attributable to out-of-school factors. Likewise, active-learning strategies like in-class peer review may only contribute to better essays in classes where students have similar writing skills and have the time to receive feedback from multiple classmates, revise their essays to accommodate feedback, and receive supplementary guidance from the teacher.

A successful RCT result can play a role in helping to warrant a prediction that the

intervention tested will work here since a positive study result indicates that the intervention *can* produce the effect at least under some set(s) of circumstances. There is still much to add, however, since the whole point is that the same intervention can perform differently in settings with different underlying social structures. So, far more research effort needs to be devoted to understanding which kinds of social/economic/cultural/material structures can afford success with an intervention and which cannot. (For example, see our discussion of Posey-Maddox's study below.) This way, decision-makers will have better resources for choosing the intervention that is best for them or for assessing and adjusting structures to accommodate new practices. Additionally, this kind of information can help educators who may want to adapt an intervention to allow for better integration with other practices or simply to capitalize on the potential it has within their setting.

Support factors are all those usually less obvious features that need to operate with the intervention for it to achieve its intended effects in the local setting. [12] Causes, including interventions, rarely work alone. They require support from various factors that are easy to overlook if we focus narrowly on isolating and measuring the effects of interventions. For example, available technology and computer literacy may be support factors for educational software. Values and norms within the school and community also count as support factors. *Derailers* are those disruptive factors that can undermine the intervention. Geography might serve as a derailer for interventions that assign substantial amounts of homework or require students to work with classmates outside of school. If many students commute from rural areas, they might not have time to do the homework and may not have access to their classmates. Even if all support factors are present, a significant and ineliminable detracting factor may count against choosing an intervention—especially if the potential derailer is a valuable school asset we would not wish to eliminate. A thriving collaborative environment that facilitates interaction and group work might be a 'derailer' relative to an intervention that mostly requires working independently, for instance.

To make plans and evaluate predictions about how effective they will be, decision-makers need to decide whether their setting is at all right for the intervention and then gauge whether the requisite support factors are in place in their setting or can be put into place with reasonable cost and effort, as well as what the chances are of derailers occurring down the line to disrupt the process. What then is needed from the research community is more theorization and a far broader body of empirical work that can help with these issues, even if the information supplied is not as certain as the RCT results that currently take center stage.

Researchers can also help by identifying the underlying causes that contributed to the problem in the study populations so that educators can consider whether their problem has the same causes as it did in the study population where the intervention worked to provide a solution. Sometimes an intervention may be an effective solution to a problem with different underlying causes, and sometimes the causes may not be relevant. But we cannot rely on that to hold generally.

Consider, for example, practices that involve ability-grouping—assigning students to classrooms or groups according to their perceived level of abilities. Creating more homogenous groups is supposed to solve the problems teachers face when their classes contain students with varying abilities and needs. Students at all levels are thought to benefit from instruction that caters to their abilities and from learning alongside peers who are similar in that respect (Colangelo et al., 2004; Gentry, 2014; Steenbergen-Hu et al., 2016). Such practices have many advocates who laud their significant positive effects on student achievement across ability groupings, while many critics argue that they are harmful to students who are socially disadvantaged (Carbonaro, 2005; Carbonaro & Gamoran, 2002; Ladd, 2012; Oakes, 2005).

Plausibly, the different outcomes stem in part from that fact that wide variation in performance among students in the same grade is a problem with different underlying causes. In places where variance grows out of natural differences among students, ability-grouping may affect

students' experiences and outcomes differently than in schools where social disadvantage significantly contributes to the variation. In the latter case, sorting mechanisms—no matter how objective they are intended to be—will likely be unreliable because social disadvantage affects both talent development and performance. Also, implicit biases may unfairly affect the process by influencing expectations and interpretations of performance. For these reasons, assessments of ability may be inaccurate and focusing on ability may actually undermine educational goals. Targeting underlying causes may not always be feasible or desirable but understanding them can inform predictions about which solution will be most effective for aiding achievement or pursuing other aims.

Many of these issues can be tackled by educational theory and research that is currently marginalized within the dominant EBE framework. Qualitative methods like case studies and ethnographies can illuminate social structures and analyze their interactions with processes or programs. They can study the causes of local problems and identify factors that may support or detract from improvement efforts.

For example, in her ethnographic case study of Morningside Elementary in Northern California, Posey-Maddox (2014) examines the role of parental involvement in an urban school. As a participant observer, she intensively collected data over the course of one school year using a range of qualitative methods including interviews with parents, teachers, and community members, observation, anonymous surveys, and analysis of relevant documents. Posey-Maddox collected further data for one year after her time at Morningside Elementary, returning for follow-up interviews and observation. Two years later, she returned for additional follow-up to assess the longer-term effects of the parental involvement programs at Morningside (Posey-Maddox, 2014, Appendix B).

Posey-Maddox uses a variety of theoretical tools and frameworks to analyze the data she

collected during her time at Morningside. She describes unintended consequences stemming from the school's reliance on parental volunteers to fill resource gaps and identifies factors that likely contributed to those outcomes, like shifting student demographics and disparities in parents' social capital. Additionally, she distinguishes between various types of parental contributions and explores ways in which each of them shaped dynamics within the school. Her analysis of the context, processes, and effects associated with parental involvement draws on social theory, including Bourdieu's theories of social power structures (e.g. Bourdieu & Passeron, 1990; Bourdieu & Nice, 1977).

In addition to providing conceptual tools to aid interpretation, using theory allows researchers like Posey-Maddox to develop plausible, albeit defeasible, explanations and explore possible implications beyond the case at hand. Her analysis of Morningside Elementary reveals factors that *could* affect interventions and identifies the circumstances in which they might do so. It thus alerts educators to potential costs and benefits of various options for designing programs, like opportunities for parental involvement. More broadly, this kind of study demonstrates how particular norms, beliefs, values, and dynamics that may be present elsewhere bear on general processes within schools.

Although these findings do not speak directly to what will happen in other cases and thus do not supply *conclusions*, they identify *premises* that are relevant to predictions about similar programs in similar contexts. Unlike RCTs, ethnographies and case studies attempt to understand which factors made a difference and how they did so. This information can help others determine whether their settings are similar in ways that affected outcomes in the study setting. Relatedly, they can offer insight into how costs and benefits were distributed across students and the factors that influenced that distribution.

For instance, Posey-Maddox attributes the negative consequences she observed to race- and

class-based inequalities within the larger community that affected the nature and outputs of parental involvement. These consequences may not occur at a school serving a racially and socioeconomically homogenous population, even if parents become involved in the same ways. What her ethnography illustrates is the potential for race and socioeconomic status to detract from efforts to fill resource gaps by involving parents in certain environments. More generally, it suggests that social structures outside of the school are relevant to the success of such programs and to their distribution of benefits and burdens. Ignoring these factors can undermine improvement efforts. Equally, assuming that they are always relevant without considering when and how they matter might lead decision-makers to dismiss interventions that could work for them. Credible theory examining relationships between education and external social structures, especially those involving social power, could help draw out the significance of these findings for other cases (e.g. Horvat & Lareau, 2003; Lareau, 2011).

To be sure, educational theory and non-experimental research have value beyond the role they could play in effectiveness predictions. Indeed, academic communities engage in this work and numerous academic journals are dedicated to publishing it. However, these enterprises usually run parallel to educational research for EBE. Their potential to contribute to the EBE model as we are suggesting remains underappreciated. As a result, information from these literatures is not prepared or mobilized for use by educational decision-makers. Intermediaries like the WWC emerged in part because proponents of EBE recognize that academic journals generally do not highlight practical implications and educational decision-makers rarely consult them (Gorard et al., 2017; Nuthall, 2004; Phillips, 2007).

Existing work from these disciplines may be useful for identifying necessary premises and finding evidence for effectiveness predictions. Applying these methods within EBE can expand its relevance. Researchers could investigate questions concerning context, social structure, support

factors, and derailers alongside RCTs. Similarly, they could study interventions that have been tested by RCTs as they are implemented in new settings or analyze failed attempts to use them. Learning about successful adaptations and failures can inform predictions and local planning.

Some researchers have undertaken such projects, but their efforts are largely directed at aiding implementation rather than prediction. Consider the study of *Success for All* (SFA) conducted by Datnow and Castellano (2000). After experimental studies found positive effects on students' literacy across multiple sites, Datnow and Castellano investigate "how teachers respond to SFA and how their beliefs, experiences, and programmatic adaptations influence implementation" (p. 777). Their findings yield multiple suggestions that can help decision-makers predict what will happen if they implement SFA in their own sites. For instance, nearly all teachers adapted the program in some way as a response to local factors or perceived deficiencies of SFA. Considering the most common adaptations, teachers' reasons for implementing them, and how they affected outcomes could help decision-makers assess, among other things, the need for and likelihood of adaptations in their own setting and evaluate their estimated impact. In these ways, their findings provide guidance for applying local knowledge and professional judgment within predictions. While some of their other findings are more useful for implementation than for prediction, designing the project with predictions in mind could lead to results that more directly support them.

When considering interventions tested by RCTs, it is useful for educators to evaluate how well the study population and setting represent their own along relevant dimensions. Information about which demographics are relevant to that particular intervention and how they affect it can be of great help to them here. Currently, the WWC provides the same demographic information for all interventions (e.g. minority status, qualification for National School Lunch Program) without specifying which are likely to affect effectiveness.

Considering only those interventions with populations and contexts that are representative

in terms of all observed characteristics will likely leave research-users with few choices. Moreover, the categories the WWC uses are too broad to be useful in many cases. For example, there are surely many differences among students who qualify for the reduced-cost school lunch program and among those with minority status. How does low socioeconomic or minority status bear on the intervention in question? Can it be expected to have a similar impact on other interventions? We cannot expect all low socioeconomic communities or households to have the same assets, support factors, and derailers. Socioeconomic status may be a better indicator of family dynamics and parental behavior in some places compared to others and those behaviors and dynamics may vary qualitatively, for instance (Furstenberg, 2011). Additionally, some interventions depend on these factors more than others. Without some understanding of why and how the results emerged, decision-makers may hastily dismiss interventions that have failed in broadly similar settings without asking whether they could have been effective for them, perhaps with some adaptation and improvement. For these reasons, research that helps educators assess representativeness along the relevant dimensions could be especially useful.

Educational decision-makers must also consider costs and benefits. Effectiveness predictions play an important role in a larger, all-things-considered decision-making process (see Brighouse et al., 2018). Educators need to know what effects *they* can expect so they can decide whether those benefits outweigh expected costs. The average treatment effect documents the difference in effects between intervention and comparison groups. Recall point (b) in the *RCT conclusion* formula. For educators, estimating the average effect they can expect requires comparing their current curriculum to what was used in the study's comparison group. If they are using a much better literacy program than the comparison group, for example, they should *prima facie* expect a smaller average effect or even a negative one. Likewise, if their program is much worse, the effect might be greater. Again, recall point (a) in the *RCT conclusion*.

Often, cost-benefit calculations at the local level concern individual students, not averages. Knowing how an intervention affected particular individuals or students with certain characteristics can usefully inform predictions about who is likely to benefit and who may bear costs. This information is especially important given persistent achievement gaps between racial and socioeconomic groups. Knowing average effects is insufficient for choosing interventions that will help (or at least not harm) underserved students.

Return to the case of ability-grouping. Meta-analyses of multiple RCTs and second-order meta-analyses consistently show positive effects on average (Steenbergen-Hu et al., 2016), leading some to recommend ability-grouping (e.g. Gentry, 2014). However, these averages do not document effects on students from particular groups. A particularly strong endorsement claims that, because "talent cuts across all demographics: ethnicity, gender, geography, and economic background," ability-grouping is particularly beneficial to talented students from disadvantaged groups (Colangelo et al., 2004, p. 7). But this is just an assumption—the data is only disaggregated by ability groups, not by social groups. Even if it were, averages for subgroups do not account for intersectionality, accuracy of sorting, different applications, or quality of experience. On average, students of color may benefit because the most advantaged among them improve dramatically while those with the lowest socioeconomic status are negatively impacted, for example.

Research that can shed light on these issues should be considered, even if its findings are less certain than RCT results. For instance, recall Posey-Maddox's observation that students from middle-class families gained greater benefits from parental involvement programs than more disadvantaged students and examines factors that likely contributed to the uneven distribution of benefits (2014, p.100-105). Although qualitative studies do not document effect sizes or precisely quantify benefits, they provide some relevant evidence for premises that concern costs and benefits. Addressing the harms stemming from educational inequalities continues to be a central goal for

education policy and practice in the U.S. Given its urgency, there is pressing need for relevant information to inform predictions concerning individual students.

While we have offered some suggestions, we are calling on the research community to seriously investigate just what can be done at a research level to help local decision-makers identify and find the facts they need to predict if an intervention is likely to work in their setting, what it would take to get it to do so, and whom it might help and whom it might harm. In this enterprise, the best should not be the enemy of the good. It is no use insisting that the information supplied be information that one can be fairly certain is correct – as in policing and reporting causal ascriptions supported by well-done RCTs—when it is not the information educators need. A far more ambitious, and riskier, program of research, theorizing, and reporting needs to be undertaken if 'evidence' is really going to help improve educational outcomes.

IV. Complementary Calls for Expanding the Research Agenda in EBE

Our proposal resembles some others that suggest diversifying research approaches and incorporating educators' local knowledge, judgment, and expertise into the decision-making process (e.g. McIntyre, 2005; Bridges & Smeyers & Smith, 2009; Smeyers & Depaepe, 2007; Bryk, 2015; Hammersley, 2015; Brown et al., 2017). But none focus on the source of the research-practice gap that we identify: unwarranted effectiveness claims. Some think general effectiveness claims can be established more quickly and reliably by supplementing evidence from research with evidence from practice. Others advise using these resources to connect the two 'communities' or 'worlds' of research and practice by translating general effectiveness claims into practical suggestions. Still others focus on improving implementation protocols. Additionally, some recent policy initiatives encourage place-based or practice-based interventions.[13] The idea is to identify promising interventions created within educational contexts as responses to problems and to figure out how to

scale them up. While some of them propose recasting the relationship between research and practice communities as bidirectional (e.g. Farley-Ripple et al., 2018), these new suggestions for connecting research and practice tend to preserve the division of labor that assigns establishing causal claims to researchers and implementation to educators. From our perspective, these suggestions get something right that will help address the gap that concerns us, though not intentionally for that reason, and are thus potentially useful within the adjusted, context-centered EBE approach we endorse.

IV.1 Research-Practice Partnerships

Research-Practice Partnerships (RPPs) are gaining momentum as a strategy for bridging the research-practice gap (Coburn & Penuel, 2016).[14] RPPs are collaborations between educators and educational researchers. They aim for research with greater relevance to practice and to improve the use of research in decision-making and practice. The driving idea is that research outcomes will be more applicable to practice if educators influence the research agenda. Educators share pressing problems with researchers who then study interventions targeting those issues. Researchers can directly help educators interpret the findings and decide how to use them in practice.

For example, Stanford University's School of Education has partnered with the San Francisco Unified School District to "help Stanford researchers produce more useful research and to help San Francisco administrators use research evidence to inform their decisions" (Wentworth et al., 2017, p. 244). One of their projects studied the outcomes of an ethnic studies course the district was piloting in some of their schools. Researchers used a regression discontinuity design to study the effects of the program on five school-year cohorts from three schools in the district. Ninth graders whose GPA was below 2.0 in the previous year were automatically enrolled in the course while students with GPA's at or above 2.0 were not. Those enrolled could opt out and others could opt in.

The study compares students just below the 2.0 threshold to students just above it because students in these two groups are taken to be similar (Dee & Penner, 2016). They found that the course positively affected GPAs and attendance for students assigned to the course, who were identified as at-risk of dropping out (Dee & Penner, 2016; Wentworth, 2016). District leaders used these findings to decide whether to implement the course throughout the district.

In this case, school administrators needed data to answer a specific question and researchers designed a study to obtain it. Researchers then helped administrators interpret the data, highlighting what it means for the decision they needed to make. While the research questions differ, this case exemplifies the typical relationship between practitioners and researchers involved in RPPs. The partnership is supposed to be mutually beneficial in that researchers obtain results with broader significance that they can publish in academic journals while practitioners get data that is directly relevant to them.

Organized this way, RPPs support the standard division of labor between educators and researchers. Practice guides research in the sense that it influences the research question. Beyond that, interaction primarily involves researchers helping to interpret results, drawing out their implications for practice. Researchers still produce information to serve as evidence in decisions about practice while educators focus on implementation.

RPPs pinpoint *relevance* of evidence as a key step in bridging the gap. We too underline the need for research to produce relevant evidence. However, RPPs generally aim for results relevant to particular *learning outcomes* whereas we urge evidence relevant to *effectiveness predictions.* Researching interventions that target local problems does not reduce the need for predicting effectiveness. In addition to aligning research projects with school districts' goals, RPPs seek generalizable results, which they suggest can be achieved using experimental or quasi-experimental methods (sometimes with the help of statistics), contrary to what we have argued. The research projects Stanford

conducts in partnership with the San Francisco Unified School District are supposed to meet "high standards of validity and generalizability" so they are relevant to other districts aiming for similar learning outcomes (Wentworth, 2016, p. 68). To serve our purposes, RPPs would need to address a wider array of questions. As it is, they are motivated by the idea that "education research does not influence policy because it takes too long to produce, is too expensive, is not applicable to a specific context of interest, and is not disseminated in a clear and direct manner" (López Turley & Stevens, 2016, p. 6S).

From our point of view, university-partnered single-school RPPs like those pioneered by four University of California campuses take a more promising approach.[15] Researchers and educators collaborate at all levels of school design and practice. Quartz, et al. (2017) describe their distinctive RPPs as "multidimensional and multilevel problem-solving ecologies" that are committed to "democratic participation" wherein "researchers and practitioners…bring their knowledge to the table and together 'build the plane while flying'" (p. 144-145). The university-partnered single-school design departs from standard RPPs by integrating researchers and practitioners. Instead of working with educators to identify problem areas and test potential solutions, help them implement best practices, or replicate an existing, evidence-based solution, researchers are familiar with available causal pathways and, in collaboration with educators, design solutions that are likely to be effective within that particular context.

For example, researchers in the Education Department at UC Berkeley partnered with Aspire Public Schools to prepare students for college. After working together for more than ten years, Aspire and UC Berkeley designed and co-founded CAL Prep, a public charter school based on local knowledge and "community-engaged scholarship."[16]Their strategies were "developed through careful and systematic study of the conditions that allowed all students to meet high academic standards" (Quartz et al., 2017, p. 144-145). While, according to Quartz, et al. (2017), these RPPs

strive to create knowledge that generalizes beyond their settings, their approach equips them to share information about support factors, underlying causes of the problems they respond to, and individual students which, we have argued, are important for local predictions and planning.

IV.2 Implementation Science

Implementation science has emerged as another response to the research-practice gap. The National Implementation Research Network (NIRN) describes it as "the science related to *implementing* [evidence-based] programs, with high fidelity, in real-world settings" (AI HUB, Module 1, n.d.) According to prominent implementation scientists Blasé (2011, 2013) and Fixsen (2009, 2010, 2013), the way to bridge the gap between research and practice is to create the infrastructure for successful implementations within education sites.-The University of North Carolina, Chapel Hill hosts the State Implementation and Scaling-up of Evidence-based Practices Center (SISEP), funded by the U.S. Department of Education. SISEP is a project within the National Implementation Research Network that creates resources for educators. They present implementation as an active, recursive process of "making it happen" instead of simply letting or helping an intervention succeed (Fixsen et al., 2009, p. 532; NIRN, 2016). Implementation scientists have developed an infrastructure that affords the capacity to implement effective interventions or innovations. It is made up of five integrated frameworks called Active Implementation Frameworks (AIFs).,

Roughly, AIFs encourage and support high-fidelity implementations of evidence-based interventions within particular local contexts (Blasé et al., 2015; Bryk, 2015; Carroll et al., 2007; Fixsen et al., 2009; Fixsen et al., 2017; Fixsen & Ogden, 2014). While implementation capacities are considered to be standard across educational settings, AIFs encourage users to take characteristics of their own sites into account at various stages. Implementation begins by exploring options using a Hexagon tool to assess the extent to which available interventions and their expected effects fit with

local needs, priorities, and existing programs (Metz & Louison, 2018).

Once they decide to implement an intervention, teams install it by making practical arrangements and "developing the knowledge, skills and abilities of teachers and administrators" through training and coaching (AI HUB, Module 1, n.d.). From there, they build toward full implementation. In part, this involves establishing and sustaining implementation drivers—core components that ensure competence among those engaging in implementation efforts, develop supports for the program, and assign leadership roles. These drivers capture "common features that exist among many successfully implemented programs and practices" (AI HUB, Module 1, n.d.). The additional frameworks concern implementation teams, which monitor implementation infrastructure and employ AIFs, and improvement cycles, which offer tools for identifying and solving problems that occur during implementation (Blasé et al., 2015).

This brief description oversimplifies AIF, but it sketches the contours well enough to contrast it with our proposal. Although it recognizes that local context can significantly impact implementation and eschews the idea that educators can simply apply research findings in practice, implementation science still adopts some of the problematic assumptions driving the dominant EBE model. In particular, it appears to assume that, in Seckinelgin's words, "the integrity of the research presented stands alone on its scientific grounds… context becomes only an issue once we move to consider policy implementation, focusing on particular target/risk groups and the way we can deliver to them what we know works" (2017, p. 132). Implementation science focuses on scaling up evidence-based programs by helping schools cultivate the capacities for successful implementation because it accepts that interventions that produce a positive outcome in a handful of good RCTs can generally be expected to do so elsewhere if implemented with fidelity, barring positive reasons to the contrary. As Fixsen (2013) puts it, the aim is to "take these good ideas that work in some places and get them to work in all places." Like other proposals to bridge the research-practice gap, it neglects

the question of whether an intervention *can* work in a specific target when implemented, whether with or without fidelity.

Using AIFs cannot ensure that an intervention can or will be effective in a particular setting. The Hexagon tool for selecting interventions directs attention to local context but focuses on whether it allows high-fidelity implementation. It does not ask whether the intervention *can* work in that setting as it did in study settings. In other words, it leaves the prediction phase out of the selection process. This is a problem because assessing the feasibility of implementation without predicting effectiveness could lead educators to choose an intervention that can be implemented in their setting but cannot work in their setting. To see this, consider a simple example: a computer program that produced positive effects for study populations when students used it in two-hour increments three times per week. Implementing the program in their own site requires educators to schedule two-hour time slots for students to use the program three times per week. Successful implementation, however, does not guarantee that it will produce positive effects there. There are many reasons that it might not: By making time for it they could inadvertently eliminate support factors for the program or activities that contribute to morale; their students might stop concentrating on it after just half an hour for any number of reasons; or, six hours per week may be too much or too little time given their students' skill sets. There is an important difference between the question of whether the local context affords a causal pathway through which the intervention can make a positive contribution and the question of whether it will allow educators to implement the intervention with fidelity.

Evaluating evidence is part of the Hexagon process, but it invites users to take general effectiveness for granted if the research meets certain criteria. It prompts users to evaluate the strength of evidence according to the number of studies that have been conducted using experimental methods with diverse populations, much like the WWC (Metz & Louison, 2018). It

does encourage users to compare their setting and population to the study and replication sites. These are factors relevant to predictions. But, recall that an RCT does not identify what population or contextual characteristics are important. Unless one replicates the whole set of post-allocation differences with exact fidelity, there is no evidence that the difference will work even granting projectability.[17]

We want to reiterate that perfect fidelity is not possible. Educational settings are so complex that reporting how interventions were implemented in a study likely leaves out details that could be relevant. More importantly, aiming for fidelity is very often not the best way to implement an intervention. Given significant differences across students and contexts, even if the target closely resembles the study along visible dimensions, producing positive effects in new settings will very likely require adjustments. Appreciating this, Nuthall (2004) observes that "the contextual details that [have] been eliminated from these studies in order to make the results generalizable are what teachers needed to know in order to … apply the results" (p. 286).

The Hexagon tool attempts to address these issues by advocating interventions with clearly defined components and a logic model or theory of change that can help educators make only 'safe' adjustments and assess progress. Although descriptions of the intervention components and implementation are likely to be incomplete, we agree that understanding key components and how they are supposed to produce outcomes would be helpful during implementation. However, prior to that they should inform predictions, which ask if an intervention *can* work in the target setting as it would be implemented there.

Notice that, just as they do not identify relevant population and contextual characteristics, RCTs do not show which components of interventions are essential to the causal process nor do they provide theories of change. Those who designed the interventions may offer theories, but they are seldom tested, and, again, RCTs cannot be used to verify them. Positive RCT results do not

show that a theory of change is accurate. Calling for this kind of information, then, signals the need for evidence from different sources to support different claims. Even if it were available, though, neither the Hexagon tool nor the other AIFs offer guidance for evaluating the kind of research and theory that could evidence these claims or speak to alternative pathways through which the intervention could make a positive contribution. Importantly, to be relevant to predictions, theories of change must account for support factors and derailers specific to the local target. While they provide some helpful information, even well-supported theories of change conceptualized by researchers or developers cannot identify these in advance. We suggest that far more effort be dedicated to providing the resources educators need to choose interventions that are likely to work for them and to make an implementation plan that is best for their setting.

IV. 3 Improvement Science, Networked Improvement Communities, and Practice-based Evidence

Another solution encourages complements to RCTs under the rubric of *improvement* science. Improvement science investigates variation in educational outcomes and devises strategies for addressing the sources of variation so interventions can work more effectively across contexts. Instead of building capacities for implementing programs with fidelity, as implementation science does, improvement science builds "capacity within the organization to understand the factors that shape improvement" and "to notice and learn from variation" (Lewis, 2015, p. 59). Whereas implementation science attempts to avoid modifications that undermine the intervention by clearly specifying its components and monitoring fidelity at the implementation site, improvement science does not recommend fidelity. Instead, it attempts to avoid modifying interventions in a way that undermines them by monitoring indicators predicted by theories of change and contextual factors that are taken to shape improvement (Lewis, 2015).

For example, Bryk (2011, 2015) proposes a collaborative effort among educators to

investigate why effects from the same intervention vary across educational settings. This would involve looking for the kinds of information we outlined in section III. Understanding what causes variation in effectiveness, he argues, allows educators to identify contextual factors that affect results. Sharing their findings widely through networked improvement communities (NIC) provides "practice-based evidence" that educators can use for better implementation (Campbell et al., 2017).

In a similar vein, Brown, with Schildkamp and Hubers (2017), suggests integrating scientific research with locally-collected quantitative and qualitative data that educators use to identify goals and inform action for reaching them. Gutiérrez and Penuel (2014) argue that establishing that an intervention 'works' requires qualitative research from educators who can provide information about how they used interventions within their settings. These proposals highlight the need for practical knowledge and action-research to supplement experimental research.

Like us, Bryk (2015) identifies an important, but neglected, distinction within EBE. He agrees that research shows only that an intervention *can* work but claims that what educators really need is "knowledge of how to actually *make it work* reliably over diverse contexts and populations" (p. 469, our emphasis). Contrast this with our claim that what educators need to know is that the intervention *will* work *here*, *in their setting*. Bryk assumes that evidence-based interventions *can* work across educational settings if the intervention or setting is properly adjusted. He advocates figuring out how to replicate results in different settings to produce "quality student outcomes reliably at scale" (2015, p. 475). While they concern more specific target settings, these suggestions still aim for general effectiveness claims that can justify predictions. We definitely agree that if there are reasonably reliable general claims about the *kinds* of populations and settings an intervention works for and how it can be used effectively in these, this is important information to secure. But we fear that there will often be too few of these kinds of reliable claims to provide much help. Looking for this sort of information is important, but it cannot replace the need for helping educators to find the

necessary information and to piece it together to make local effectiveness predictions as we are suggesting.

If we treat the information from practice as evidencing general effectiveness claims, then these proposals are unlikely to improve significantly on the existing model. Using the argument theory, we can more easily see what practice-based evidence can be evidence for and how it can be used responsibly in policy deliberations. Instead of (or in addition to) using practice-based evidence to improve implementation or gain evidence of general effectiveness, educators should use it for their predictions. Reporting on variation in outcomes and trying to determine which local variables made the difference can be very useful for learning what conditions affect the effectiveness of particular interventions. But we cannot use knowledge from practice to neutralize sources of variability, making interventions work reliably across contexts. Nor can we rely on the accumulated knowledge as an inductive base that warrants conclusions about other targets on its own. A large and varied evidence base cannot warrant a prediction without further premises supporting the claim that the results will travel. Even then, induction provides weak evidence compared to premises that speak to the local structure and the support factors needed there.

The upshot is that collecting and disseminating results and practical knowledge through NICs and other knowledge mobilization networks can help to close the gap between research and practice if educators use the evidence to make predictions rather than continuing to abide by the standard division of labor. Whether it comes from educators or researchers, predictions require information about the conditions that affect effectiveness no matter what intervention is under consideration. Of course, not all information obtained this way will be equally reliable. Responding to the suggestion that individual craft knowledge or knowledge from action-research can be coordinated and compiled to provide a useful, evidence-based body of professional knowledge that educators can use to improve their practice, Nuthall (2004) points out that "what is going on in a

classroom that leads to student learning is more complex and difficult to disentangle than a teacher has time to record, analyze, and interpret" (p. 292). Additionally, educators' interpretations may be influenced by biases, leading them to pass on inaccurate, or potentially harmful, information.

Causal ascriptions and explanations can be true or false and the reasoning offered in support can be better or worse. Well-warranted predictions require premises that are supported by strong evidence. The relevant information is difficult to assess rigorously. As such, to make good use of alternative research types and reports from practice we need mechanisms for evaluating the claims.

It seems that the WWC and similar databases could be well-positioned to evaluate and disseminate these resources. Currently, the WWC includes only original research from RCTs or quasi-experimental studies. Much of what we and others are recommending is *secondary* research, including conceptualization and theorizing. For example, reports about successful and failed attempts to use an intervention in practice do not currently qualify as 'evidence' by their standards. Reports about individual cases and various alternative forms of educational research can be found on blogs, websites, or in academic journals. These are ineffective channels both because they are inconsistently accessed and because some lack credibility. Even if NICs provide better avenues for sharing results, knowledge transmitted that way could be misleading. Databases could evaluate these materials and organize them according to research type. They could also communicate how particular resources might be useful and how they should *not* be used in predictions. The last thing we want to do is bombard educators with more information without guidance for how to use it alongside experimental research and their own local knowledge in deliberations about policy and practice.

Conclusion

We attribute the persistent gap between what worked in research and what works in practice in part to lack of support for the effectiveness planning and prediction central to the standard EBE

model. We distinguish between three kinds of causal claims: causal ascriptions, general effectiveness claims, and effectiveness predictions. We argue that, at best, educational RCTs evidence causal ascriptions, which, without further assumptions, are irrelevant to general effectiveness claims and effectiveness predictions. Because general effectiveness claims are not essential for predictions and are difficult to establish, we propose a serious rethinking of the EBE model to figure out how better to produce evidence and theory relevant to effectiveness predictions directly. Recognizing the sort of considerations that are necessary to support local predictions suggests a far broader, context-centered research agenda. Additionally, materials for decision-makers should highlight local planning and prediction as an indispensable step.

Examining other recent strategies for addressing the research-practice gap, we find that they can be helpful for facilitating our proposal but are not, on their own, enough to bridge the gap. If the planning and prediction phase remains invisible, educators invited to collaborate with researchers are unlikely to request the information most relevant to their predictions. Educators can surely influence research agendas by identifying widespread problems. Research that investigates those problems will be relevant to practice in a topical sense. But, bridging the gap between research and practice requires more than topically relevant research or more detailed plans for implementation and adaptation—it requires research that is relevant to local effectiveness predictions.

Notes

We are grateful to Adrian Simpson for helpful feedback on an early draft of this paper and to Harry Brighouse for his thoughtful input. Funding for this project was provided to Nancy Cartwright and Kathryn Joyce by the Center for Ethics & Education. For Cartwright, this material is based upon research supported by the National Science Foundation under grant no: 1632471 and the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 667526 K4U). It is acknowledged that the content of this work reflects only the authors' views and that the ERC is not responsible for any use that may be made of the information it contains.

1. For example, Brown, Schildkamp & Hubers (2017) suggest integrating local knowledge with rigorous evidence of effectiveness, Bryk (2015) calls for networks wherein educators share information they learn in the course of implementing evidence-based interventions, and Jinfa Cai, et al., (2017) urge researchers to "offer information on effective ways to implement" effective interventions (p. 345). In Section IV, we consider these proposals in more detail.
2. See Cartwright and Hardie (2012) for the general structure of this frame; for essential components see Rothman (1976) on 'support' factors and Bechtel & Abrahamson (2005) on underlying structure.
3. EBE, in its current form, may have arisen from disputed positivist doctrines, but that doesn't mean that all versions of EBE owe what justification they have to those doctrines. Thus, instead of targeting the doctrines taken to motivate EBE, as many have done, we directly address problems present within EBE. See Kvernbekk (2016) for an excellent discussion distinguishing the *necessary* attributes of EBE as a basic concept from the attributes attached to particular conceptions of EBE, like the dominant 'what works' model.
4. For a general discussion of context-centered versus intervention-centered approaches, see Cartwright (2019), or the extended discussion of context centering in the area of international HIV Aids policies in Seckinelgin (2017), much of which applies to education.
5. We cannot account for what makes theory credible with any precision in the abstract. However, academic standards and expertise on the part of theorists in relevant fields could be used to assess theory.
6. This is the average of the 'individual treatment effects' of the individuals in the study population, i.e., how much difference the intervention would make to the individual supposing all other causes of the outcome were the same. Amazingly, RCTs allow for an estimate of the average of individual treatment effects even though we cannot measure these counterfactual values themselves. For more on this see Rubin (1974).
7. More technically, if randomization, blinding, and other post-random-assignment policing succeed in ensuring the intervention is probabilistically independent in the mean from the net effect of all 'other' causes of the outcome, the difference in means between intervention and control groups will be an *unbiased estimate* of the average intervention effect in the study population – which can be far from a correct (or 'precise') estimate. For more on unbiasedness versus precision in RCTs, see Deaton & Cartwright (2018).
8. Simpson (2017) notices that 'strong evidence of positive effects' is commonly misinterpreted to mean 'evidence of strong positive effects.'
9. As Frances Bacon taught in 1620, "Induction by simple enumeration is puerile." (*NO* I:105)
10. As with the estimates of average treatment effect in an RCT on a study population, the estimate should be better as the size of the sample increases, but the problem never goes away entirely.
11. School effectiveness research has been subject to some criticism, especially by advocates of using research evidence to inform education policy. See for example Coe & Fitz-Gibbon (1998), Brown, et al. (1997), Goldstein & Woodhouse (2000). We do not rely on this research in a substantial way, nor are we claiming that it should inform policy. Rather, we take the enterprise and its general observations to indicate that schools, although not wholly independent of one another and the broader education system, differ in ways that *may* bear on the performance of educational interventions in some cases. Thus, we should not assume differences between study and target schools are negligible when assessing generalizability. For more on assumptions about representativeness see Joyce (2019).
12. The distinction between structural features on the one hand and support factors and derailers on the other is not a hard and fast one. But it is useful to separate factors that are deeply entrenched and difficult for the relevant educators to change – which Munro, et al. (2016) label 'structural' – from ones that educators can more readily change or substitute for (like by an after-school

homework club where students from distant homes can work together) – which they label 'support factors' and 'derailers'.
13. The Every Student Succeeds Act of 2015 created programs to support innovations developed locally by educators. For example, the Education Innovation and Research Program (ESSA, sec. 4611) provides funding for "evidence-based, field-initiated innovations" that can be scaled-up to help more students.
14. The U.S. Institute for Education Sciences has introduced programs that encourage RPPs.
15. These are the University of California, Los Angeles, UC Berkeley, UC San Diego, and UC Davis.
16. See the Center for Educational Partnerships website: cep.berkeley.edu/cal-prep.
17. As Adrian Simpson noted in commenting on a draft of this paper (personal communication, April 11, 2018).

References

AI HUB (n.d). Module 1: An overview of active implementation frameworks. Accessed September 19, 2018. https://implementation.fpg.unc.edu/module-1.

Archibald, T. (2015). "They just know": The epistemological politics of "evidence-based" non-formal education. *Evaluation and Program Planning*, *48*, 137–148. https://doi.org/10.1016/j.evalprogplan.2014.08.001

Bacon, F. (2000). *The new organon.* (L. Jardine & M. Silverthorne, Eds.). New York: Cambridge University Press.

Bechtel, W. & Abrahamsen, A. (2005). Explanation: A mechanist alternative, *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 36: 421-441.

Becker, B. E. & Luthar, S. S. (2002). Social-emotional factors affecting achievement outcomes among disadvantaged students: Closing the achievement gap. *Educational Psychologist*, *37*(4), 197–214. https://doi.org/10.1207/S15326985EP3704_1

Berliner, D. C. (2002). Comment: Educational research: The hardest science of all. *Educational*

Biesta, G. (2007). Why "what works" won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory*, *57*(1), 1–22. https://doi.org/10.1111/j.1741-5446.2006.00241.x

Biesta, G. J. J. (2010). Why 'what works' still won't work: From evidence-based education to value-based education. *Studies in Philosophy and Education*, *29*(5), 491–503. https://doi.org/10.1007/s11217-010-9191-x

Blasé, K.A., Fixsen, D. and M. Duda. (2011). *Implementation science: Building the bridge between science and practice*. Invited Presentation to the Institute of Education Sciences, Dept. of Education, Washington, DC.

Blasé, K. & Fixsen, D. (2013). Core intervention components: Identifying and operationalizing what makes programs work. ASPE Research Brief, Office of the Assistant Secretary for Planning and Evaluation, Office of Human Services Policy, U.S. Department of Health and Human Services.

Borman, G. D. (2009). The use of randomized trials to inform education policy. In *Handbook of Education Policy Research*. Routledge. https://doi.org/10.4324/9780203880968.ch11

Bourdieu, P., & Passeron, J. C. (1990). *Reproduction in education, society, and culture*. Newbury Park, Calif: Sage in association with Theory, Culture & Society, Dept. of Administrative and Social Studies, Teesside Polytechnic.

Bourdieu, P., & Nice, R. (1977). *Outline of a theory of practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511812507

Bridges, D., Smeyers, P., & Smith, R. (Eds.). (2009). *Evidence-based education policy: what evidence? what basis? whose policy?* Malden, MA: Wiley-Blackwell.

Bridges, D., & Watts, M. (2008). Educational research and policy: Epistemological considerations. *Journal of Philosophy of Education*, *42*, 41–62. https://doi.org/10.1111/j.1467-9752.2008.00628.x

Bridges, D. (2017). *Philosophy in educational research*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-49212-4

Brighouse, H., Ladd, H. F., Loeb, S., & Swift, A. (2018). *Educational goods: values, evidence, and decision making*. Chicago: The University of Chicago Press.

Brighouse, H., & Schouten, G. (2011). Understanding the context for existing reform and research proposals. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? rising inequality, schools, and children's life chances*. Chicago: Russell Sage Foundation; Spencer Foundation.

Brown, S., Duffield, J., & Riddell, S. (1997). School effectiveness research: the policy maker's tool for school improvement? In N. Bennett, A. Harris, & M. Preedy (Eds.), *Organizational effectiveness and improvement in education* (pp. 138–146). Philadelphia: Open University Press.

Brown, C. (2014). *Making evidence matter: a new perspective for evidence-informed policy making in education*. Institute of Education Press.

Brown, C., Schildkamp, K., & Hubers, M. D. (2017). Combining the best of two worlds: a conceptual proposal for evidence-informed school improvement. *Educational Research*, *59*(2), 154–172. https://doi.org/10.1080/00131881.2017.1304327

Bryk, A. S. (2015). 2014 AERA distinguished lecture: Accelerating how we learn to improve. *Educational Researcher*, *44*(9), 467–477. https://doi.org/10.3102/0013189X15621543

Bryk, A. S., Gomez, L. M., & Grunow, A. (2011). Getting ideas into action: Building networked improvement communities in education. In *Frontiers in Sociology of Education* (pp. 127– 162). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1576-9_7

Campbell, C., Pollock, K., Briscoe, P., Carr-Harris, S., & Tuters, S. (2017). Developing a knowledge network for applied education research to mobilise evidence in and for educational practice. *Educational Research*, *59*(2), 209–227. https://doi.org/10.1080/00131881.2017.1310364

Carbonaro, W. J., & Gamoran, A. (2002). The Production of achievement inequality in high school English. *American Educational Research Journal*, *39*(4), 801–827. https://doi.org/10.3102/00028312039004801

Carbonaro, W. (2005). Tracking, students' effort, and academic achievement. *Sociology of Education*, *78*(1), 27–49. https://doi.org/10.1177/003804070507800102

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*(1). https://doi.org/10.1186/1748-5908-2-40.

Cartwright, N. (2010). What are randomised controlled trials good for? *Philosophical Studies*, *147*(1), 59–70. https://doi.org/10.1007/s11098-009-9450-2

Cartwright, N. (2013). Evidence, Argument and Prediction. In V. Karakostas & D. Dieks (Eds.), *EPSA11 Perspectives and Foundational Problems in Philosophy of Science* (pp. 3–17). https://doi.org/10.1007/978-3-319-01306-0_1

Cartwright, N. (2017). Single Case Causes: What Is Evidence and Why. In H.-K. Chao & J. Reiss (Eds.), *Philosophy of Science in Practice* (pp. 11–24). https://doi.org/10.1007/978-3-319-45532-7_2

Cartwright, N. (2019). *Nature, the artful modeler: Lectures on laws, science, how nature arranges the world and how we can arrange it better*. Chicago, IL: Open Court.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better*. Oxford; New York: Oxford University Press.

Coburn, C. E., & Penuel, W. R. (2016). Research–practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, *45*(1), 48–54. https://doi.org/10.3102/0013189X16631750

Coburn, C. E., & Stein, M. K. (Eds.). (2010). *Research and practice in education: building alliances, bridging the divide*. Lanham, Md: Rowman & Littlefield Publishers.

Coe, R., & Fitz-Gibbon, C. T. (1998). School effectiveness research: criticisms and recommendations. *Oxford Review of Education*, *24*(4), 421–438. https://doi.org/10.1080/0305498980240401

Colangelo, N., Assouline, S. G., & Gross, M. U. M. (2004). *A nation deceived: How schools hold back America's brightest students* (Vol. 1). Templeton Foundation. Retrieved from http://www.accelerationinstitute.org/Nation_Deceived/ND_v1.pdf

Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using randomised controlled trials in education* (1st edition). Thousand Oaks, CA: SAGE Publications.

Crotty, M. (1998). *The foundations of social research: meaning and perspective in the research process*. London; Thousand Oaks, Calif: Sage Publications.

Curto, V. E., Fryer Jr., R. G., & Howard, M. L. (2011). It may not take a village: Increasing achievement among the poor. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? rising inequality, schools, and children's life chances*. Russell Sage Foundation; Spencer Foundation.

Datnow, A., & Castellano, M. (2000). Teachers' responses to Success for All: How beliefs, experiences, and adaptations shape implementation. *American Educational Research Journal*, *37*(3), 775–799. https://doi.org/10.3102/00028312037003775

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Dee, T., & Penner, E. (2016). The causal effects of cultural relevance: Evidence from an ethnic studies curriculum. CEPA Working Paper No. 16-01. Retrieved from Stanford Center for Educational Policy Analysis: http://cepa.stanford.edu/wp16-01.

Duncan, G. J., & Murnane, R. J. (Eds.). (2011). *Whither opportunity? rising inequality, schools, and children's life chances*. New York: Chicago: Russell Sage Foundation; Spencer Foundation.

Duncan, G. J., & Murnane, R. J. (2014). *Restoring opportunity: the crisis of inequality and the challenge for American education*. Cambridge, Massachusetts: Harvard Education Press.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" education research. *Educational Researcher*, *32*(7), 31–38.

Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher*, *47*(4), 235–245. https://doi.org/10.3102/0013189X18761042

Fixsen, D. L., Blasé, K. A., Naoom, S. F., & Wallace, F. (2009). Core implementation components. *Research on Social Work Practice*, *19*(5), 531–540. https://doi.org/10.1177/1049731509335549

Fixsen, D. L. & Blasé, K.A. (2013). An overview of scaling-up and active implementation. [video] State Implementation and Scaling-up Evidence-Based Practices Center homepage. Retrieved from http://sisep.fpg.unc.edu/.

Fixsen, D.L., Blasé, K., Duda, M., Naoom, S. and M. Van Dyke. (2010). Sustainability of evidence-based programs in education. *Journal of Evidence-Based Practices for Schools*, 11(1), 30-46.

Fixsen, D., Blasé, K., Metz, A., & Dyke, M. V. (2013). Statewide implementation of evidence-based programs. *Exceptional Children*, *79*(2), 213–230. https://doi.org/10.1177/001440291307900206

Freiberg, H. J. (Ed.). (1999). *School climate: measuring, improving, and sustaining healthy learning environments*. Philadelphia: Falmer Press.

Furstenberg, F.F. (2011). The challenges of finding causal links between family educational practices and schooling outcomes. In G.J. Duncan & R.J. Murnane (Eds.), *Whither opportunity? rising inequality, schools, and children's life chances*. Chicago: Russell Sage Foundation; Spencer Foundation.

Gates Foundation (n.d.) K-12 education strategy overview. [webpage]. Retrieved from https://gatesfoundation.org/What-We-Do/US-Program/K-12-Education.

Gentry, M. L. (2014). *Total school cluster grouping & differentiation: a comprehensive, research-based plan for raising student achievement and improving teacher practices* (Second edition). Waco, Texas: Prufrock Press.

Goldstein, H., & Woodhouse, G. (2000). School effectiveness research and educational policy. *Oxford Review of Education*, *26*(3–4), 353–363. https://doi.org/10.1080/713688547

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education: the promises, opportunities and problems of trials in education*. Routledge.

Gutiérrez, K. D., & Penuel, W. R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher*, *43*(1), 19–23. https://doi.org/10.3102/0013189X13520289

Halsall, R., Carter, K., Curley, M., & Perry, K. (1998). School improvement: the case for supported teacher research. *Research Papers in Education*, *13*(2), 161–182. https://doi.org/10.1080/0267152980130204

Hammersley, M. (2002). *Educational research, policymaking and practice*. London; Thousand Oaks, Calif: P. Chapman.

Hammersley, M. (2013). *The myth of research-based policy & practice*. Los Angeles: SAGE.

Hargreaves, D. H. (1995). School culture, school effectiveness and school improvement. *School Effectiveness and School Improvement*, *6*(1), 23–46. https://doi.org/10.1080/0924345950060102

Horvat, E. M., Weininger, E. B., & Lareau, A. (2003). From social ties to social capital: Class differences in the relations between schools and parent networks. *American Educational Research Journal*, *40*(2), 319–351. https://doi.org/10.3102/00028312040002319

Howe, K. R. (2009). Positivist dogmas, rhetoric, and the education science question. *Educational Researcher*, *38*(6), 428–440. https://doi.org/10.3102/0013189X09342003

Jinfa Cai, Anne Morris, Charles Hohensee, Stephen Hwang, Victoria Robison, & James Hiebert. (2017). Making classroom implementation an integral part of research. *Journal for Research in Mathematics Education*, *48*(4), 342. https://doi.org/10.5951/jresematheduc.48.4.0342

Joyce, K. E. (2019). The key role of representativeness in evidence-based education. *Educational Research and Evaluation*, 1–20. https://doi.org/10.1080/13803611.2019.1617989

Joyce, K. E., & Cartwright, N. (2018). Meeting our standards for educational justice: Doing our best with the evidence. *Theory and Research in Education*, *16*(1), 3–22. https://doi.org/10.1177/1477878518756565

Kvernbekk, T. (2016). *Evidence-based practice in education: functions of evidence and causal presuppositions*. New York: Routledge, Taylor & Francis Group.

Ladd, H. F. (2012). Education and poverty: Confronting the evidence. *Journal of Policy Analysis and Management*, *31*(2), 203–227. https://doi.org/10.1002/pam.21615

Lareau, A. (2011). *Unequal childhoods: class, race, and family life* (2nd ed.) Berkeley: University of California Press.

Levin, B. (2013). To know is not enough: research knowledge and its use. *Review of Education*, *1*(1), 2–31. https://doi.org/10.1002/rev3.3001
Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher*, *44*(1), 54–61. https://doi.org/10.3102/0013189X15570388
López Turley, R. N., & Stevens, C. (2015). Lessons from a school district–university research Partnership: The Houston Education Research Consortium. *Educational Evaluation and Policy Analysis*, *37*(1_suppl), 6S-15S. https://doi.org/10.3102/0162373715576074
MacNeil, A. J., Prater, D. L., & Busch, S. (2009). The effects of school culture and climate on student achievement. *International Journal of Leadership in Education*, *12*(1), 73–84. https://doi.org/10.1080/13603120701576241
McIntyre, D. (2005). Bridging the gap between research and practice. *Cambridge Journal of Education*, *35*(3), 357–382. https://doi.org/10.1080/03057640500319065
McLeod, J. D., & Kaiser, K. (2004). Childhood emotional and behavioral problems and educational attainment. *American Sociological Review*, *69*(5), 636–658. https://doi.org/10.1177/000312240406900502
Metz, A., & Louison, L. (2018). The hexagon tool: Exploring context. *Chapel Hill, NC: National* https://implementation.fpg.unc.edu/sites/implementation.fpg.unc.edu/files/imce/documents/NIRN-HexagonDiscussionandAnalysisTool2018-FINAL.pdf
Munro, E., Cartwright, N., Hardie, J., & Montuschi, E. (2016). *Improving Child Safety: Deliberation, Judgment and Empirical Research* (ISSN 2053-2660). Retrieved from https://www.dur.ac.uk/chess/
Nelson, J., & Campbell, C. (2017). Evidence-informed practice in education: Meanings and applications. *Educational Research*, *59*(2), 127–135. https://doi.org/10.1080/00131881.2017.1314115
NIRN. (2016). Active implementation practice and science. *National Implementation Research Network*.
Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*(4), 647–670. https://doi.org/10.1086/378858
Nuthall, G. (2004). Relating classroom teaching to student learning: A critical analysis of why research has failed to bridge the theory-practice gap. *Harvard Educational Review*, *74*(3), 273–306. https://doi.org/10.17763/haer.74.3.e08k1276713824u5
Nutley, S. M., Walter, I., & Davies, H. T. O. (2007). *Using evidence: how research can inform public services*. Bristol, U.K: Policy Press.
Oakes, J. (2005). *Keeping track: how schools structure inequality* (2. ed). New Haven, Conn.: Yale Univ. Press.
Phillips D. C. (2007). Adding complexity: Philosophical perspectives on the relationship between evidence and policy. *Yearbook of the National Society for the Study of Education*, *106*(1), 376–402. https://doi.org/10.1111/j.1744-7984.2007.00110.x
Posey-Maddox, L. (2014). *When middle-class parents choose urban schools: Class, race, and the challenge of equity in public education*. Chicago: The University of Chicago Press.
Quartz, K. H., Weinstein, R. S., Kaufman, G., Levine, H., Mehan, H., Pollock, M., … Worrell, F. C. (2017). University-partnered new school designs: Fertile ground for research–practice partnerships. *Educational Researcher*, *46*(3), 143–146. https://doi.org/10.3102/0013189X17703947
Reiss, J. (2013). What's wrong with our theories of evidence? *Theoria, 29(2)*, 283-306. DOI: http://dx.doi.org/10.1387/theoria.10782
Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): a state-of-the-art review. *School Effectiveness and School Improvement*, *25*(2), 197–230.

https://doi.org/10.1080/09243453.2014.885450
Rothman, K. J. (1976). Causes. *American Journal of Epidemiology*, *104 (6),* 587–92.
Rowan, B. (2011). Intervening to improve the educational outcomes of students in poverty: Lessons from recent work in high-poverty schools. In G. J. Duncan & R. J. Murnane (Eds.), *Whither opportunity? rising inequality, schools, and children's life chances* (pp. 523–538). Sage Foundation; Spencer Foundation.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350
Sarason, S. B. (1990). *The predictable failure of educational reform: can we change course before it's too late?* (1st ed). San Francisco: Jossey-Bass.
Sarason, S. B. (1996). *Revisiting "The culture of the school and the problem of change."* New York: Teachers College Press.
Schoen, L. T., & Teddlie, C. (2008). A new model of school culture: a response to a call for conceptual clarity. *School Effectiveness and School Improvement*, *19*(2), 129–153. https://doi.org/10.1080/09243450802095278
Scriven, M. (1994). The final synthesis. *Evaluation Practice, 15(3)*, 367–382. http://dx.doi.org/10.1016/0886-1633(94)90031-0
Seckinelgin, H. (2017). *The politics of global aids: institutionalization of solidarity, exclusion of context*. New York, NY: Springer Berlin Heidelberg.
Shavelson, R. J., Towne, L. and National Research Council (U.S.), eds. (2002). *Scientific research in education*. Washington, DC: National Academy Press.
Simpson, A. (2017). The misdirection of public policy: comparing and combining standardised effect sizes. *Journal of Education Policy*, *32*(4), 450–466. https://doi.org/10.1080/02680939.2017.1280183
Slavin, R. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 31(7): 15–21.
Smeyers, P., & Depaepe, M. (Eds.). (2007). *Educational research: Why 'what works' doesn't work*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-5308-5
Spillane, J.P. and Miele, D.B. (2007). Evidence in practice: A framing of the terrain. *Yearbook of the National Society for the Study of Education,* 106(1): 46-73.
Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement: Findings of two second-order meta-analyses. *Review of Educational Research*, *86*(4), 849–899. https://doi.org/10.3102/0034654316675417
Tough, P. (2013). *How children succeed: grit, curiosity, and the hidden power of character.*
Tseng, V., & Nutley, S. (2014). Building the infrastructure to improve the use and usefulness of research in education. In K. S. Finnigan & A. J. Daly (Eds.), *Using research evidence in education* (pp. 163–175). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-04690-7_11
Wentworth, L., Carranza, R., & Stipek, D. (2016). A university and district partnership closes the research-to-classroom gap. *Phi Delta Kappan*, *97*(8), 66–69. https://doi.org/10.1177/0031721716647024
Wentworth, L., Mazzeo, C., & Connolly, F. (2017). Research practice partnerships: a strategy for promoting evidence-based decision-making in education. *Educational Research*, *59*(2), 241–255. https://doi.org/10.1080/07391102.2017.1314108

What Works Clearinghouse. (n.d.). Find what works homepage. [website]. Retrieved from https://ies.ed.gov/ncee/wwc/FWW.

What Works Clearinghouse (2016a). Interventions summary for Read 180. [webpage] Retrieved from https://ies.ed.gov/ncee/wwc/EvidenceSnapshot/665.

What Works Clearinghouse. (2016b). *Intervention report for Read 180*. Retrieved from https://ies.ed.gov/ncee/wwc/Intervention/742.

What Works Clearinghouse. (2017a). *Standards handbook 4.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

What Works Clearinghouse. (2017b). *Procedures handbook 4.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf